DOCUMENT RESUME

ED 400 294 TM 025 570

AUTHOR	Scheuneman, Janice Dowd; Slaughter, Carole
TITLE	Issues of Test Bias, Item Bias, and Group Differences and What To Do While Waiting for the Answers.
PUB DATE	May 91
NOTE	33p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ethnic Groups; *Group Membership; *Item Bias; *Minority Groups; Psychometrics; Sex Differences; Test Use; *Test Validity
IDENTIFIERS	Item Bias Detection

ABSTRACT

A number of explanations have been offered for the differences in test performance among various population subgroups. This paper begins with a discussion of these explanations including the psychometric explanation that group differences are due to bias in the test. An overview of bias research argues that results to date are inconclusive. A theory of bias is introduced that provides a definition of bias and a framework that explains why the issues are so difficult to resolve. Bias is defined as the systematic over- or underestimation of the true abilities of a group of examinees formed according to some demographic variable such as sex or ethnicity. The framework also provides a connection between test bias and item bias. The concept of item bias is then distinguished from that of differential item functioning (DIF). DIF research is described as promising in many regards, but with major areas of uncertainty in the interpretation of its results. It has generally supported the reliability and validity of standardized tests for minority groups, but as long as research is based on hypothetical scenarios instead of solid, research-based theory, the question of whether test bias accounts for some portion of the observed differences between groups is likely to remain unanswered. Practical guidance is offered for those who must make important decisions about individuals without knowing the answers about test bias. (Contains 26 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 400 294

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Janice Dowd Scheuneman
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Issues of Test Bias, Item Bias, and Group Differences

and What to do While Waiting for the Answers

Janice Dowd Scheuneman, Carole Slaughter

Educational Testing Service

May, 1991

BEST COPY AVAILABLE

TM025570

The ideas presented in this paper are those of the authors and are not to be construed as having been endorsed or approved by the Educational Testing Service.

Abstract

A number of explanations have been offered for the differences in test performance among various population subgroups. This paper begins with a discussion of these explanations including the psychometric explanation: Group differences are due to bias in the test. An overview of the bias research follows which argues that the results to date are inconclusive. A theory of bias is then introduced which provides a definition of bias and a framework that enables us to explain why the issues are so difficult to resolve. This framework also provides the connection between test bias and item bias. The concept of item bias is then distinguished from that of differential item functioning (DIF). DIF research is described as promising in many regards, but also having major areas of uncertainty in the interpretation of results. Finally, the paper offers practical guidance for those in the field who must make important decisions about individuals without knowing the answers about test bias.

One of the major areas of controversy in measurement today is test bias. Testing is increasingly being used for important decisions such as licensing, certification, and admissions to colleges, universities, and graduate and professional schools. Many of the tests used for these purposes show differences in mean scores between groups defined by sex, race, ethnicity or socioeconomic status, raising concern that individuals from groups with lower scores may be unfairly denied access to employment or educational opportunities.

Moreover, a growing proportion of the U.S. population's workforce in the future will come from what are now labelled "minority groups," and from women. This growth will occur at a time when this country will simultaneously experience an increasing demand for highly trained, technologically capable personnel. Today, American business spends significant amounts to train workers, including teaching them reading, writing, and other basic skills. Identifying talent and identifying it early are becoming increasingly more to both the public and private sectors. These factors are likely to provide additional pressure to think more carefully about which tests are used, what scores mean, and how tests are used. The national need will require that we do so.

Thus testing is an area where the real-life stakes are high, and debate about bias in testing is complicated by the use of terms that have different meanings to many of the participants in the discussion. In psychometrics, bias is a statistical term with a quite explicit and neutral meaning, but it has connotations in general usage which are largely negative. Even psychometricians, however, have no generally agreed-upon definition of bias as that term is applied to tests and critics untrained in measurement vary even more in their perceptions of the meaning of the term when it is applied by measurement specialists.

Nevertheless, the debate rages as if all participants were talking about the same concept.

The purpose of this paper is to attempt to clarify some of the issues surrounding test bias. In the first section, we will discuss group differences in test scores and the various hypotheses that have traditionally been offered to explain these differences. In the following sections, we will explore the meaning of bias from the perspective of measurement research and theory. Finally, the last section provides suggestions for the test user who must continue to make important decisions while the questions about bias remain unanswered. The ideas expressed in these sections are our own, but they are consistent with widely accepted views and research findings. We believe that they provide a basis for understanding and integrating different aspects of the controversy, as well as illustrating why the issues of bias are so difficult to resolve.

Group Differences

In studying group differences, the practice has typically been to compare average test scores of "groups" that have been constituted by using the traditional sociological variables of ethnicity, race, sex, or socioeconomic status. Most analyses show a difference in mean scores of one to one and one-half standard deviations between minority and non-minority test-takers, with the frequent exception of Asian-Americans. Differences exist, too, between females and males, particularly in mathematics and the sciences. Older adults will also typically score less well than younger adults on many types of tests.

One difficulty that arises from making such comparisons is the possibility that group membership may somehow be perceived as causing a lower score. People from groups identified as typically obtaining lower scores on tests are often concerned that perceptions of their groups' "ability" will adversely affect their individual opportunities for educational and employment advancement. They are concerned that the general public will fail to understand that many individuals in their group obtain very high scores and that the meaning of any one test score is limited.

To fail to discuss group differences in test scores, however, leaves everyone, including those policy and decision-makers who believe in issues of equity, without the knowledge necessary to put test performance in proper perspective. Before suggesting ways in which such data can be properly evaluated by decision-makers, it is important to discuss how such differences are typically explained.

A review of various attempts to explain test score differences among groups indicates that such explanations fall into five broad categories: historical, cultural, biological, educational, and psychometric. These categories are not mutually exclusive. In practice a set of explanations posited to account for specific score differences may include a combination of two or more of these categories. A brief description of the five areas and the types of explanations that fit within each is provided below. The intent is not to determine the merits of the various explanations, but to provide a structure for discussion of ways in which test score differences are explained by various individuals and interest groups.

Historical explanations generally center around past practices, such as de jure and de facto segregation which, in this country, has had the effect of producing unequal access to a variety of facilities and experiences that

enhance an individual's knowledge base. For example, the access to libraries, museums, and other institutions of learning in many places was limited or non-existent under the de jure segregation of African-Americans in the South until relatively recently in our history. In analyzing score differences between African-Americans and Whites on graduate and professional school tests, an investigator seeking to explain these differences might note first that the African-American test-taking population is relatively older. The investigator might then note that many of these test-takers are working to overcome the deleterious effects of segregation on possibilities for learning, i.e., that one reason for test score differences is historical deprivation.

Cultural explanations generally center on behaviors, language issues, styles of learning, and ways of being that appear, at least on the surface, to have a significant relationship to performance on traditional multiple-choice standardized tests. For example, when attempting to explain the often superior test performance of students from some Asian cultures, the authoritative structure of the family in these cultures is often cited. Another type of explanation that would be cultural in nature would be the suggestion that the culture of some American Indian tribes is extremely non-competitive and that this non-competitive environment in which young people are immersed has a profound impact on their ability to take tests designed to rank order individuals. Many explanations that center on socio-economic status would also be classified here, mainly because many of them tend to be explanations that focus on the concept of a "culture of poverty," rather than simply on the fact that a person is poor and may therefore not do well on tests.

Biological explanations of test score differences are ones that attempt to explain the differences in terms of some innate feature of individuals or groups, or environmental factors within the womb. An explanation of the tendency of females to perform more poorly than males on mathematics tests that suggests that "girls just can't do math" implies some innate deficiency. Current discussions about how the babies of crack addicts are likely to perform on a variety of measures seek to make a link between the fetus and later test performance. Biological explanations, especially those that center on assumptions of the innate nature of groups, tend to be perceived as particularly inflammatory. This is, at least partially, because they seem to contain an implicit assumption that achieving equality of performance, no matter what the interventions, is impossible.

Differences in educational experiences as explanations include essentially three types of discussions: (a) the number and quality of courses taken by specific groups, as well as the amount of time on task; (b) the quality of teachers and teaching given to various groups; and (c) the motivation of students as it relates to their experiences within the educational environment. An example of an educational explanation of why Puerto Rican students do not score as well as White students on college entrance tests is one that points to the smaller number of academic courses taken by the Puerto Rican students. Also educational in nature is the explanation that students from urban high schools perform less well on achievement tests because they are more likely to have had teachers who were not trained in the subjects they must teach.

Psychometric explanations are those which point to the tests themselves, which many believe fail to adequately measure the knowledge and abilities of the many test takers. That is, "bias" in the test is assumed to be the explanation for most or all of the observed group differences. Issues of test speededness, test center conditions, and test-taking skills fall here as well. In many instances, the operationalization of these concerns from the perspective of the "public" are quite differert from the meaning of the terms when used by psychometricians and others immersed in the world of measurement.

From the "lay" point of view, the inclusion of material that is, for example, esoteric or seemingly irrelevant is sufficient evidence of bias in itself. Psychometricians, however, generally mean something quite different. For many measurement specialists, a test in sociology that never mentions females, Hispanics, or poor people may be considered unbiased if it passes certain statistical criteria. The result is that psychometricians and the "public"--be they test-takers, policy-makers, or others--often talk past each other. Thus, it is difficult for any resolution of the conflicts to occur. It is important here to point out that measurement is a relatively new "science" and new methodologies are constantly being developed that permit us to look at existing data in different ways and to analyze emerging data differently. It is also important for psychometricians to keep in mind that absence of proof is not proof of absence.

In the following section of this paper, the focus will be on the psychometric issues--that is, those concerning test bias. These issues will be discussed within the context of a theoretical framework that defines bias in abstract terms. The major threads of bias research will then be related to

that framework. The purpose of this discussion is to demonstrate some of the reasons why issues of bias have proved so difficult to resolve even within the measurement community.

Bias in Test Scores

"Bias," as a statistical term, indicates the presence of systematic error in a statistic when estimating an unknown quantity. In the context of testing, the quantity we are trying to estimate is the "true" score on a test, the score that reflects the real, but unobservable (latent) level of ability within an individual. If it were possible to administer a test to an individual a large number of times without any effects of practice, fatigue, boredom, etc., we would expect the scores on this test to vary, but the average of these scores would, in theory, be equal the individual's "true" score on the test. Similarly, if a large sample of people from a population subgroup is tested, the mean of the observed scores of the group is expected to closely approximate the true score mean of the individuals in that group. In the terms of measurement theory, then, a test may be said to be "biased" for a group of examinees if the average of the observed scores systematically underestimates the true score mean of that group (Jensen, 1980; Petersen, 1980; Scheuneman, 1984).

Because the true scores cannot be observed, a number of problems immediately arise in attempting to determine if bias in fact exists. The practice has been for a researcher to develop a plausible scenario for a particular instance of test use together with the expected outcome if the test is biased. One such scenario is that if a test is underestimating the abilities of members of a minority group, we might expect that the grades they earn in college will be higher than predicted from the test scores. This is a testable hypothesis that could be evaluated using a simple regression design (Cleary, 1968).

A number of such studies have been done, but they have generally failed to demonstrate that this effect is occurring. In fact, in such studies, grades for minority groups are often found to be lower than predicted (Jensen, 1980; Hunter, Schmidt & Rauschenberger, 1984).

While many observers have thereby concluded that tests are not biased in these situations, others have pointed out that there are problems with this method of study. For example, another scenario that has been suggested is that people earning equivalent grades in college should have obtained equivalent admissions scores on an unbiased test. It is a peculiarity of the regression method, however, that this scenario and the one discussed in the previous paragraph are mutually exclusive. That is, using real test and criterion settings, a test found to be unbiased in one of these scenarios, must be biased in the other (Darlington, 1971). More recent evaluations of the regression procedures by Linn (1984) suggest that under most realistic circumstances, the logical outcome suggested in the first scenario is unlikely to be observed, even if test scores were in fact biased against the minority group, unless the degree of bias was very large. Others have pointed out the effects on the results of either implicit or explicit pre-selection on another variable (Linn, 1983; Swinton, 1981) or problems arising from a criterion measure that may also be biased (Flaugher, 1978; Williams, Mosby, & Hinson, 1978).

Still another possible scenario is that if scores are being underestimated for a group, the test must be measuring something different for that group. For instance, a test may be incidentally measuring an additional domain of knowledge or skill beyond that which it is intended to measure. If this domain is one which is more likely to be known by the White majority test taker and is not a valid component of the ability the test is intended to

measure, the result may be to underestimate the scores of minority examinees on the intended domain. Hence, another strategy for determining if a test is biased is to analyze what the test measures for different groups, most often using factor analytic procedures. Examples are studies by Gutkin and Reynolds (1981), Johnston and Bolen (1984), Reschly (1978), and Sandoval (1982).

The results of factor analytic studies have generally supported the similarity of factor structure for minority and White examinees. Hence, the tests studied do appear to be measuring largely the same thing for different population groups. Some indications may be found among the results, however, that suggest the factors are not identical (Scheuneman, 1981). That is, while the tests measure largely the same thing for different groups of examinees, other less salient components of performance may have slightly different effects on scores. If so, whether the effects of such components are valid for the purposes of the test or whether the effect results only in an increase in random error, rather than a systematic underestimation of true scores, is unknown.

Is it possible for tests to be both biased and valid? Many investigators would argue that a biased test cannot be valid. We would argue that validity is necessary, but not sufficient for an unbiased test (Scheuneman, 1984). Consider the following scenario. Suppose a study comparing the heights of men and women is being conducted. The men and women are being measured in different sites. Due to a mixup in instructions, however, men are measured in centimeters and women in inches. The results would show that men on the average are taller, as expected, but the difference would appear larger, and possibly of more significance, than it really is. Alternately, suppose that those measuring the women quickly noted that their height varied widely according to the height of the heels on their shoes and had all women remove

their shoes to be measured, although they had not been instructed to do so. Those measuring the men had them keep their shoes on. Again, the result would be that the men on the average were taller than the women, but again the difference between the groups would appear larger than it really is.

These two examples represent the problems that would arise, respectively, due to differences in the size of the units of the scales of measurement and differences in the origins of these scales. Notice that, in both cases, the measurements were reliable and valid measures of height, correlating as expected with other variables, and properly rank ordering individuals within each group. If a cut-off were to be used, however, such that only people attaining a certain height could be admitted to a certain program, the women would be underrepresented in relation to the number who should qualify. The height of the women would be underestimated in relation to the height of the men and the measurement could be said to be biased against them.

The possibility that measurement units might be different for two groups can be evaluated using confirmatory factor analytic methods. Although that was not the purpose of the studies, such methods were used by Rock and Werts (1979) and Rock, Werts, and Grandy (1980). These studies confirmed the hypothesis of equivalent factor structures for Black and White examinees, thereby supporting the validity of the tests being investigated, but the hypothesis of equal scale units was rejected for four of the five test scales. The effect sizes were quite small, however, and the practical significance of these differences on the test scores is unknown.

The examples presented here do not provide an exhaustive accounting of the types of research on test bias that have been pursued, but the underlying pattern is the same in all such studies. A scenario is envisioned in which the effects of bias, if it exists, might be observed and an appropriate experimental design or data analysis for capturing such effects are developed.

(An area of investigation sometimes also referred to as test bias research actually is concerned with fairness of test use and not with bias as it is defined here. See Cole, 1981; Jaeger, 1976; or Petersen, 1980 for discussions of this topic.)

In summary, the evidence from this work strongly supports the validity of the tests for minority groups for the same purposes as they are used for Whites, given the examinees have sufficient English language competence. Further, the results seem to indicate that gross amounts of bias in the scores of minority group members do not exist (Cole, 1981; Jensen, 1980). On the other hand, there are sufficient problems and anomalies in the results of these studies for some researchers to believe that the evidence falls short of establishing that no bias exists (Scheuneman, 1987; Shepard, 1987).

To date, a body of research has yet to emerge that all observers can agree demonstrates that the scores of minority examinees are or are not biased. This lack of certainty leaves people free to accept or reject the various findings according to which of these agree with their individual "biases" concerning what they believe to be true. Given that the true ability or skill we are trying to measure is unobservable and that the stakes of testing are so high, this situation is likely to remain unchanged for some time to come. Significant progress is not likely as long as we are working from hypothetical scenarios rather than from solid, research-based theory concerning group differences in characteristic modes of learning and cognitive processing as well as theory on how these differences interact with test materials.

Toward a Theory of Test Bias

Bias, in this statistical sense, can be detected in scores only at the group level. Although bias clearly can affect the score of an individual examinee, any one test score is always just an approximation of his or her

true ability. Further, if a test, which is essentially valid, tended to be biased against examinees from a particular group, this would not become apparent with regard to legitimate uses of that test unless these examinees were to be compared with a second group of examinees for whom the test is not biased or is biased to a different degree. (Consider the above example of men's and women's heights.)

Suppose then that for any one group we were able to determine the amount of bias resulting from each item in a test--that is, the over- or under-estimation of the true ability of the examinee group contributed by each of the items. Theoretically, items could vary considerably in this regard. Some items could be essentially unbiased, with a bias quantity near to zero; others could over estimate ability, still others might under estimate it, some to a considerable extent. We could then determine a mean and standard deviation of these bias quantities across the different items in the test. If the mean were near zero, the test score would be essentially unbiased, even if the standard deviation, and hence the degree of bias in some items, were quite large. On the other hand, if all items consistently over or under estimated the ability of a group of examinees, so that the mean of the bias quantities was significantly different from zero, the score would be biased to that degree, even if the standard deviation were very small.

If we define a biased item to be one with a particularly large bias quantity, we can see from the above that the presence of biased items is neither a necessary nor a sufficient condition for a biased test. The presence of biased items will result in a larger standard deviation of the theoretical bias quantities, but may or may not much affect the mean depending on the other items on the test. If the mean is held constant, it can readily be demonstrated that increasing the standard deviation of the bias quantities,

even to a considerable extent, has little effect on the resultant scores (Scheuneman, 1981). Test bias is the result of the mean of the individual item contributions and it can exist even if the amount of bias does not vary much across items and no particular items have an exceptionally high or low bias value.

Let us assume that such bias quantities exist for each group of interest. Thus the net effect of the bias in group comparisons would be the difference between the bias means for two groups. Typically we think of bias resulting from factors that have adverse effects on the performance of, for example, minority examinees. The impact of bias on performance differences between Whites and minorities may equally well arise from a test that is largely unbiased for most minority group members but biased in favor of White examinees, whose scores therefore tend to overestimate their ability. While the impact on scores is the same either way, the causes of the bias for these different possibilities--bias against certain minorities or in favor of the White majority--are likely to be quite different. This is, therefore, an important distinction to keep in mind when seeking to understand how bias might arise.

The degree of over or under estimation of ability in an item might vary according to a number of different factors. Such factors might be traced to (a) properties of the test in general or of the individual items, (b) personal characteristics that might tend to be more (or less) prevalent among examinees from the group of concern than among those from the group to whom they are being compared, or (c) an interaction between the item and examinee characteristics. Some such factors will affect the test as a whole, and hence most likely the mean of the bias quantities, while others will affect individual items, causing the variation in bias across items to increase.

Examples of possible contributors to bias arising from the test might include differences in the adequacy of instructions for persons from different groups, especially for novel material or tasks; the item format or mode of presentation of the item task; differential attractiveness of the item key or distractors in multiple-choice tests; test length, which may result in differential speededness; and items that are differentially difficult according to the strategy used to arrive at a correct response.

Personal characteristics could also result in differential group performance, if these characteristics were distributed differently in the groups being compared. Such characteristics might include various personality attributes, cognitive styles, interest or motivation within the testing situation, or negative feelings related to school or assessment settings. While these personal characteristics are likely to affect the test as a whole, individual items may also be differentially susceptible to the effects of personal attributes. Different backgrounds and experiences could cause certain strategies to be used more often in one group than another. The manifest content of some items could elicit more interest and possibly more careful attention to the item task demands. Material in an item might be offensive and arouse negative emotions or include implicit assumptions that are so in conflict with the examinee's world view or perspective that the intent of the item is misread or misunderstood.

Another possible source of bias might be group differences in test-wiseness skills, which will often be unrelated to the knowledge, skills, or abilities being measured. Test wiseness probably develops as test taking skills are increased through experience to the point of becoming automatic, that is, requiring little active thinking from the test taker. Once this has

occurred, the examinee is likely to have the time and the cognitive resources, during the process of taking a test, to attend to nuances of items and to search for any unintentional cues that may have been left by the test developer. Experience alone is clearly not sufficient for the development of test wiseness, however. While some reasoning ability is needed, a more important factor may be a perception that the test is a challenge and that the testing situation is basically unthreatening. Such perceptions seem most likely to result when previous experiences with education and assessment have usually been positive. Unfortunately, group differences are also likely to exist in the extent to which positive experiences have previously been linked with testing.

Item Bias and Differential Item Functioning

Historically, item bias procedures were developed in the mid-1970's to meet a need to screen items for possible bias during the test construction process. Bias detection procedures available in the literature at that time all required comparing the test score with some outside criterion measure of ability, which could not be done until after a new test had been administered. Moreover, if procedures could be applied during the test construction process, problematic items could be modified or deleted from the item pool prior to administration of the test for real-life purposes. The procedures developed were called item bias procedures to distinguish them from the criterion-related methods, which were often called test bias procedures.

Fundamentally, the item bias procedures were designed to detect possible bias in instances where direct comparisons of the performance of two groups were inappropriate because the groups were also known to differ in factors related to the development of the relevant skills and abilities and to the

acquisition of relevant knowledge. Some critics of testing have asserted that the performance of two groups in an unbiased test should be identical, but this argument assumes the true score means of the two groups are the same, an assumption that is often not warranted in the light of other evidence.

Over the years, a number of different procedures for the detection of possible item bias have been advanced and evaluated. The methods now most often recommended define an unbiased test item as one where examinees of equal ability have equal probability of getting the item correct, regardless of group membership. Ability, for this purpose is defined by the observed score on the test or test section or by an estimated true score based on item response theory (IRT) methods. Notice that IRT true scores are estimated from performance on the test and hence are not independent estimates of "true" ability that might be used to determine if the test score is biased. (Reviews of the item bias procedures are provided by Hills, 1989; Rudner, Getson & Knight, 1980; and Scheuneman & Bleistein, 1989.)

Because the item bias methods determine ability based on test performance, they have been criticized for assuming the test as a whole is unbiased. Although some researchers may make this assumption in interpreting their results, this is not, in fact, a necessary assumption for the methods. The only necessary assumption for the methods is that the test is a valid and reliable measure for all groups being compared--an assumption that is generally supported by the research discussed in earlier sections of this paper. This implies that, within each group, the test properly discriminates between those of high ability and those of low ability and does a reasonably satisfactory job of rank ordering individuals on that ability dimension. The statistics yielded by these procedures will then do a fairly good job of

sorting out those items on which the lower scoring group does relatively well compared to the higher scoring group and those on which it does particularly poorly. That is, we can identify those items which depart furthest from the expectations that arise if equal probabilities for equal scores are assumed.

The problem with the item bias methods is that one group can do particularly well or particularly poorly on a test item when compared to another group for reasons other than bias in the item. That is, any systematic difference in the way two groups respond to an item can be reflected in a significant bias statistic. Differences in performance between two groups of examinees may thus be related to differences in life experiences and cultural values as well as to differences in previous exposure to the material in the item or to opportunity to learn. If such differences result in genuine differences in the level of the knowledge, skill or ability being measured by the test, which are reflected in performance differences on an item, such differences would not be considered bias. Recognition that the item bias procedures detect more than just bias led to the introduction of the term, differential item functioning (DIF), which is thought to be more descriptive of what is found when using these procedures.

A critical part of a DIF analysis, therefore, is to determine the sources of the observed difference. For example, if a group is found to do less well on problems involving fractions on a mathematics achievement test, we would probably look to the preparation received by these groups to explain the outcome rather than to look for an explanation in the test items. In some instances, however, features of an item may differentially affect the capability of examinees from these groups to demonstrate the appropriate knowledge, skill or ability required to respond correctly to the item. We would probably conclude that these features result in bias in the items.

For example, if a math item for elementary school students requires examinees to estimate the weight of a football, it may favor boys over girls. Unfortunately, in practice, this distinction between differences that are validly related to the purpose of the test and those which are not is very difficult to make. The reasons why the observed differences may have arisen often cannot be determined, and hence the judgment of whether the DIF constitutes bias cannot readily be formed.

In practice, therefore, the DIF research rapidly enters an area of uncertainty, just as test bias research discussed earlier does, though in a different way. In both instances, highly exact quantitative methods can be brought to bear on the problems of bias in testing, but the conclusions are much less than clear cut. Again, real progress will probably not occur without a solid base of both knowledge and theory concerning relevant group differences. DIF studies differ from test bias studies, however, in that the DIF statistics will often reflect some of the group differences of interest in this regard and may be one of the best sources available for the short run for the acquisition of new knowledge about group differences.

Changing the Impact of Test Scores

Understanding that a variety of explanations for test score differences exist is important in helping to analyze the current dialogue about testing in American education. Understanding how intractable the psychometric problems are in evaluating possible bias in tests is important in realizing that such bias, if it exists, cannot be readily eliminated. However, this understanding does little in helping a decision-maker, faced with a range of test scores from a diverse pool of applicants, to determine what he or she can or should do about selection or placement decisions.

The practical day-to-day reality of an administrator, a teacher, a guidance counselor, or an admissions officer requires that he or she be concerned about such questions as: Will overall test scores begin to decline as minorities become a larger proportion of our young population? If so, should my institution reduce enrollments to maintain "quality?" Since women seem to get better grades than men and since often grades are what standardized tests attempt to predict, should I continue to use these tests? How can I know when I have a female or minority candidate who can be successful in my program if the mean scores for those groups are lower? Should I choose only students who have scores in the same range for my program? Only by acknowledging the strengths and limitations of current standardized testing, can the practitioner make an informed judgment about who gets what, when, and how.

Standardized test scores provide a relatively inexpensive and time-efficient way to compare the performance of students from a variety of backgrounds and geographic areas on a common set of material. This assumes, of course, that the tests have adequate construct validity, i.e. measure essentially the same skills and abilities for all population groups. It also assumes that they have high reliability, i.e. that a student taking the same or similar test will get, within some defined range, a similar score each time. For purposes of this discussion, acknowledging the caveats found elsewhere in this article, let us assume the test we are working with meets these criteria. How can existing test instruments be used to identify a broad range of talented individuals who may succeed in the program or profession they wish to pursue. And what directions might prove useful for test-makers to pursue in helping practitioners make good decisions.

Evaluate the Test Selection Process

Many missed opportunities for talent identification can be attributed to the selection of an inappropriate test. Tests are selected by a variety of individuals for a variety of reasons. Legislatures want evidence of "educational outcomes" provided through low-cost assessment procedures that are easy on the bureaucracy. Faculty members are often ambivalent or divided in their views. The same faculty member may change views depending upon the situation. Some want tests that will identify a "talented tenth" for them to teach, but are concerned when a difficult test given at the time of a student's exit from a program means they may be held accountable for poor performance.

Accountability, as well as issues of turf, often play as large a role in test selection as need for diagnosis, placement, etc. In some institutions standardized tests are selected by one office, scored by another, and used by a third office for remediation or acceleration. This scattershot approach to using tests rarely maximizes the institution's ability to be helpful to the individual test-taker. Those who will use the test results need to be involved in the selection of the test. When appropriate, an assessment of how the proposed test matches the curriculum should be made. It is folly to use a test to evaluate performance in a particular curriculum that does not assess the major areas of that curriculum. The flip side of this situation is failing to monitor the content of tests your students may have to take to enter college, graduate or professional school, thus leaving them to navigate such tests as best they can. This is not to say one should teach to a test. It is simply recognizing that students you may train will sooner or later be competing in a bigger pond.

To evaluate a test for use within an institution, it is necessary to determine what information is needed, what assessment instruments are available, and whether there is, in fact, a match. Verification should be made that a test has been reviewed for fairness (it never hurts to do an in-house analysis as well). Data on test performance by race, sex, and socio-economic group should also be obtained. This latter information is critical because if students who constitute a large proportion of the relevant population score poorly, it may be self-defeating institutionally to operationalize an objective of, for instance, having niñety-five percent of the test-takers answer ninety out of one hundred questions correctly within six months. This is not to suggest that this could not be a goal, but rather to indicate that in light of the data, a careful look at objectives would be critical if the test described was used.

Train Test Users

In our society, we do not let individuals drive our streets without a permit or license; one cannot become a doctor without rigorous training, and yet thousands of individuals in their roles as directors of admissions, faculty members, etc. spend some of their time using test scores to help make decisions about individuals' lives without ever being required to read about the test and its uses or take a course in statistics, or tests and measurement. Not surprisingly, this can lead to some interesting, albeit uninformed (or misinformed), interpretations of test score meaning among even the most dedicated individuals. In addition, staff turnover and rotation of responsibilities mean that training must be quick, easily available, and affordable to educational institutions. This is not to suggest that no one

but measurement experts should be involved in decision-making about individuals when assessment instruments are being used. Not only would that be unrealistic and impractical, but it would fail to recognize that the judgments being made about admissions, certification, and diagnosis are more than psychometric in nature.

Consider, however, ensuring that all decision-makers have been provided with available information about the test being used, that it has been discussed in a group setting, and that questions have been answered. Test makers, too, often provide a wealth of information about their products, and some provide training staff to work with institutions that request assistance in understanding the purpose and appropriate uses of a particular test. Too often administrators and faculty are unaware of where and to whom to turn for help. For every uninformed decision-maker at an institution, some candidate's life chances may be affected. Understanding the limitations and the strengths in assessment then is crucial.

Use Multiple Criteria

It cannot be said too often that most standardized tests should not be used as the sole criterion for a decision, such admittance into a program or institution. This is true not only because intuition tells us that a person can have a bad day, or be unfamiliar with particular test item formats, etc., but because most tests were not designed to be used that way. While it is much easier to tell an applicant that he or she was not admitted because of a specific test score than to try to describe the intricate and often complex process by which admissions decisions are made, the latter is more helpful to the candidate and certainly more honest.

Examine Within Group Scores

The very nature of a standardized test means that one is comparing abilities of individuals from different backgrounds and regions of the country. Often the assessment instrument is designed to rank test-takers based on the obtained scores. What sometimes happens, however, is that every institution wants the "best" female or minority students, regardless of whether they are in the "best" departments, or the "best" institutions. This is where a look at within group scores may prove useful in some instances.

For example, a political science department, which is a strong department, wants to enroll some Black students at the graduate level (they have had one Black student in the past six years) and they are willing to provide two fellowships for this purpose. Because the department wants to bring in "superior" students, they decide, after some discussion, that the fellowship recipients must have a 650 on the verbal and a 650 on the quantitative section of the test to be eligible for competition. After two months pass and they receive no applicants, the chairman asks how they might better tap the pool. According to an analysis of 1986-87 GRE test-takers, only 3.6 percent of 9324 Black examinees had a 650 or better on the verbal section and 1.7 percent had a 650 or better on the quantitative section of the test (Educational Testing Service, 1988). The analysis does not show how many students with a 650 quantitative score also had a 650 verbal score. Assuming that half of the high quantitative group overlaps with the high verbal group, the maximum available pool would be 0.85 percent (half of 1.7) or about 80 individuals. The same source tells us that only 137 Black students during that year indicated plans to study political science or government. Even assuming that people with such high quantitative scores would be as likely to select political science as a physical science, engineering or mathematics, 0.85 percent of this group is only about one person.

Without belaboring the point by delineating other possible factors such as the number of graduate departments in political science, the department may clearly wish to rethink its reliance on the proposed score. The pool, for example, might be significantly increased by looking at the top thirty percent of all Black students taking the test or by appropriately using multiple criteria. The department could also consider awarding the fellowships to the "best" two students in the existing applicant pool.

By carefully examining how a student's test score fits within his or her group, decision-makers can use the information to make current admissions decisions while concomitantly searching for ways to increase the size and "quality" of the available applicant pool. If departments or universities wait until parity is reached between groups on scores and use scores as the most important criterion, the wait could well be fifty years or more and result in many lost generations of individuals who could make significant societal contributions.

Prepare Test-Takers

Even the best assessment instrument is only as good as the knowledge base it draws upon and the preparedness of those who take it. Too often individual test-takers, out of fear, naivete, or lack of the knowledge of where or how to start, assume there is nothing that they need to do to prepare for a standardized test. This assumption is rarely true. Before taking a test, students need to attempt practice questions; know the timing, directions, question types, and materials that should be brought to the test. They also need to understand penalties of guessing, if any; as well as the purpose of the test. Interviews with students over the past ten years continue to indicate that significant numbers of minority students do not have access to, or do not take opportunities to engage in significant test preparation.

If we really believe that we need to broaden our definition of talent and also become better at identifying it, it is incumbent upon those who support student success to ensure that individuals learn what they need to know about test-taking and that they learn it early in life. Making sure that materials are available, setting up test preparation opportunities, and providing the student with a feeling of empowerment in the potential testing situation can be helpful.

Summary and Conclusions

A number of explanations have been offered for the persistent differences between groups on various types of tests. In this paper, we have focused on one of these explanations: the possible effects of test bias. Bias was defined as the systematic over or under estimation of the true abilities of a group of examinees formed according to some demographic variable such as sex or ethnicity. Because true ability is unobservable, however, the detection of bias must be indirect. The basic research designs have been developed by formulating a hypothetical scenario in which bias might be functioning and devising an experiment that would detect bias in this instance. Although the research has generally supported the reliability and validity of standardized tests for minority groups and little evidence of bias has emerged, sufficient problems remain for many researchers to be skeptical about this apparent lack of bias in tests. One of the most important of these problems is that different scenarios can lead to mutually contradictory results. As long as the research is based on hypothetical scenarios rather than solid, research-based theory, the question of whether test bias accounts for some portion of the observed differences between groups is likely to remain unanswered.

Different DIF procedures, designed to detect bias in items, do tend to identify the same items in reference to a particular group. They also identify items, however, on which group differences exist that are valid with respect to the purposes of the test. The problem of determining whether a given item has a high DIF value because of such valid differences or because of bias in the item turns out to be quite a difficult one. While some items can clearly be identified as valid or not, the majority call for judgments to be made on the basis of vague or unspecified criteria. Moreover, the presence or absence of biased items in a test does not constitute evidence of bias or lack of bias in test scores. DIF studies can yield useful information to contribute to the knowledge base needed for the development of theory about bias, but bring us no closer to telling us if the scores we must use in decision situations are biased or not.

Perhaps among the most difficult tasks that decision-makers in education face are selection, diagnosis, and placement decisions. In many ways these decisions have the longest-lasting effects on the opportunity structure of the individual test-taker. Since measurement experts are the first to admit that tests are useful, but not perfect, practitioners would do well to take heed. They need to understand what tests can and cannot do and to review constantly the role of assessment instruments within their institution in the creation and elimination of barriers to opportunity. Those committed to enhancing the national talent pool can do so, even using tests that may be biased, if these tests are used carefully and appropriately.

As the demographics change and as institutions confront increasingly diverse demands on their resources, it will become a significant part of the national agenda to choose talent well. Interestingly, this need for better talent identification will provide a window of opportunity to more equitably address the concerns of Black, Hispanic, and female Americans. Many bright and capable individuals who can and want to make a contribution to the success of this country may thus be identified and their talents put to use. The challenge for assessment is to help in this process in a way which encourages policies and practices of inclusion, rather than exclusion.

References

Cleary, T. A. (1968). Test Bias: Prediction of grades of Negro and White students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Cole, N. S. (1981). Bias in testing. American Psychologist, 36, 1067-1077.

Darlington, R. B. (1971). Another look at "cultural fairness." Journal of Educational Measurement, 8, 71-82.

Educational Testing Service. (1988, June). A summary of data collected from Graduate Record Examinations test takers during 1986-1987 (Data Summary Report #12). Princeton, N.J.: Author.

Flaugher, R. L. (1978). The many definitions of test bias. American Psychologist, 33, 671-679.

Gutkin, T. B., & Reynolds, C. R. (1981). Factorial similarity of the WISC-R for White and Black children from the standardization sample. Journal of Educational Psychology, 73, 227-231.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8(4), 5-11.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. R. Reynolds & R. T. Brown (Eds), Perspectives on bias in mental testing. New York: Plenum.

Jaeger, R. M. (Ed.) (1976). On bias in selection [Special issue]. Journal of Educational Measurement, 13, 1-99.

Jensen, A. R. (1980). Bias in mental testing. New York: Free Press.

Johnston, W. T., & Bolen, R. M. (1984), A comparison of the factor structures of the WISC-R for Blacks and Whites. Psychology in the Schools, 21, 42-44.

Linn, R. L. (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected seamples. Journal of Educational Measurement, 20, 1-15.

Linn, R. L. (1984). Selection bias: Multiple meanings. Journal of Educational Measurement, 21, 33-47.

Petersen, N. S. (1980). Bias in the selection rule--bias in the test. In L. J. Th. van der Kamp, W. F. Langerak, and D. N. M. de Gruijter (Eds.), Psychometrics for educational debates. London: John Wiley.

Reschly, D. J. (1978). WISC-R factor structures among Anglos, Blacks, Chicanos, and Native-American Papagos. Journal of Consulting and Clinical Psychology, 3, 417-422.

Rock, D. A., & Werts, C. E. (April, 1979). Construct validity of the SAT across populations: An empirical confirmatory study (RR-79-2). Princeton, N.J.: Educational Testing Service.

Rock, D. A., Werts, C. E. & Grandy, J. (March, 1980). Construct validity of the GRE across populations: An empirical confirmatory study. Princeton, N.J.: Educational Testing Service.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. Journal of Educational Statistics, 17, 1-10.

Sandoval, J. (1982). The WISC-R factorial validity for minority groups and Spearman's hypothesis. Journal of School Psychology, 20, 198, 204.

Scheuneman, J. D. (1981). A new look at bias in aptitude tests. In P. Merrifield (Ed.), Measuring human abilities (New Directions in Testing and Measurement, No. 12). San Francisco: Jossey Bass.

Scheuneman, J. D. (1984). A theoretical framework for the exploration of causes and effects of bias in testing. Educational Psychologist, 19, 219-225.

Scheuneman, J. D. (1987). An argument opposing Jensen on test bias: The pscyhological aspects. In S. Modgil & C. Modgil (Eds.), Arthur Jensen: Consensus and Controversy. London: Falmer Press.

Scheuneman, J. D. & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2, 255-275.

Shepard, L. A. (1987). The case for bias in tests of achievement and scholastic aptitude. In S. Modgil & C. Modgil (Eds.), Arthur Jensen: Consensus and Controversy. London: Falmer Press.

Swinton, S. S. (March 1981). Predictive bias in graduate admissions tests (RR-81-24). Princeton, NJ: Educational Testing Service.

Williams, R. L., Mosby, D. & Hinson, V. (1978). Critical issues in achievement testing of children from diverse ethnic backgrounds. In M. J. Wargo & D. R. Green (Eds.), Achievement Testing of Disadvantages and Minority Students for Educational Program Evaluation. Monterey, Cal.: CTB/McGraw-Hill.